2015

An Analysis of the Impact of Playout Delay Adjustments Introduced by VoIP Jitter Buffers on Listening Speech Quality

Peter Počta
University of Žilina, Slovakia.

Hugh Melvin
National University of Ireland, Galway

Andrew Hines
Technological University Dublin, andrew.hines@tudublin.ie

Recommended Citation

Počta, P., Melvin, H. & Hines, A., (2015) An Analysis of the Impact of Playout Delay Adjustments Introduced by VoIP Jitter Buffers on Listening Speech Quality, *ACTA ACUSTICA UNITED WITH ACUSTICA,* Vol. 101 (2015) 616 – 631. doi:10.3813/AAA.918857

Acta Acustica united with Acustica
Vol. 101 (2015) 616 – 631

DOI 10.3813/AAA.918857

An Analysis of the Impact of Playout Delay Adjustments introduced by VoIP Jitter Buffers on Listening Speech Quality

Peter Počta[1)], Hugh Melvin[2)], Andrew Hines[3)]

[1)] Dept. of Telecommunications and Multimedia, FEE, University of Žilina, Univerzitná 1, 01026, Žilina, Slovakia. peter.pocta@fel.uniza.sk

[2)] Discipline of Information Technology, College of Engineering & Informatics, National University of Ireland, Galway, Ireland. hugh.melvin@nuigalway.ie

[3)] Sigmedia, Trinity College Dublin, Ireland. andrew.hines@tcd.ie

Summary

This paper investigates the impact of frequent and small playout delay adjustments (time-shifting) of 30 ms or less introduced to silence periods by Voice over IP (VoIP) jitter buffer strategies on listening quality perceived by the end user. In particular, the quality impact is assessed using both a subjective method (quality scores obtained from subjective listening test) and an objective method based on perceptual modelling. Two different objective methods are used, PESQ (Perceptual Evaluation of Speech Quality, ITU-T Recommendation P.862) and POLQA (Perceptual Objective Listening Quality Assessment, ITU-T Recommendation P.863). Moreover, the relative accuracy of both objective models is assessed by comparing their predictions with subjective assessments. The results show that the impact of the investigated playout delay adjustments on subjective listening quality scores is negligible. On the other hand, a significant impact is reported for objective listening quality scores predicted by the PESQ model i.e. the PESQ model fails to correctly predict quality scores for this kind of degradation. Finally, the POLQA model is shown to perform significantly better than PESQ. We conclude the paper by identifying further related research that arises from this study.

PACS no. 43.71.Gv, 43.72.Kb

1. Introduction

The default best-effort Internet presents significant challenges for delay-sensitive applications such as VoIP. To cope with non-determinism, VoIP applications employ receiver playout strategies that adapt to network conditions. Such strategies can be categorised as either per-talkspurt or per-packet. The former take advantage of silence periods within natural speech and adjust such silences to track network conditions, thus preserving the integrity of talkspurts. This approach thus minimises delay at the expense of silence period adjustments and some potential late packet loss. Examples of this approach include [1, 2]. Per-packet strategies are different in that adjustments are made both during silence periods and during talkspurts by scaling of packets, a technique also known as time-warping. This approach is more responsive to short network delay changes in that the per-talkspurt approach can only adapt during recognised silences even though the timescale of many delay spikes may be less than that of a talkspurt. The main disadvantage of this approach is the degradation caused by the scaling of speech packets. Examples of the latter approach are described in [3, 4] and such techniques can be found in popular VoIP applications such as GoogleTalk and Skype. Other research has attempted to optimise buffer size and in particular, the trade off between late packet loss and delay based on customised objective models [5, 6, 7, 8] Finally, previous research by one of the authors has proposed a hybrid playout strategy that utilises synchronised time in order to implement an informed fixed delay playout whenever possible thus minimising the need for playout adjustments whilst minimising late packet loss impairments. It reverts to an adaptive approach when delays become excessive. Details of this approach can be found in [9]. In this research, we focus on applications that deploy per-talkspurt strategies, which are commonly found in current telecommunication networks.

Comparative performance analysis of the various per-talkspurt playout strategies has to date largely focused on metrics such as average delay and extent of late packet loss. We have found little research to date that has thoroughly and specifically examined the precise impact of multiple and frequent silence period adjustments, characteristic of such adaptive playout strategies on speech quality. Although both Ramjee *et al.* [1] and Moon *et al.* [2]

Received 04 October 2013,
accepted 03 December 2014.

cite Montgomery [10] in claiming that such distortion does not have a noticeable effect, the latter which was published in 1983 does not provide any evidence in this regard. All three simply qualify their assertion regarding the impact of silence period distortion by stating that small adjustments are not noticeable. On the other hand, research by Hoene *et al.* [6] has shown that playout delay adjustments during active speech have significant impacts on subjective listening speech quality, but tests did not assess adjustments during silences. Hoene *et al.* also validated the use of the PESQ model to predict the impact of adjustments during active speech. In subsequent research by Hoene *et al.* [5], PESQ was used to estimate the impact of single large adjustments during both silences and active speech, and regarding the former, shows how adjustments of up to approximately 320 ms are deemed not noticeable. Finally, research using subjective listening tests by Voran [11], suggests that very large adjustments (430 ms) are noticeable and then examines the impact of various general impairments, but not specifically silence period adjustments.

All of the above tests, both subjective and objective, address listening quality only. Other research has examined the broader issue of conversational quality which includes the interactive nature of voice communications. For example, Lee *et al.* [12] suggest that in a wider context, playout delays typical of jitter buffer strategies can, when considered at both ends of a VoIP session, have an effect in a conversational environment and thus impact speech quality. They propose a time-scaling approach that whilst impacting marginally on listening quality, minimises or eliminates the need for jitter buffer delays, thus minimising any impact on conversational quality. However, their testing approach is based solely on listening-only tests. The impact of playout adjustments on conversational speech quality was also raised by Gong *et al.* [13]. They discuss the ITU-T E-model which takes into account end-to-end delays, and thus goes some way towards examining conversational quality. Their analysis of the impact of delay on conversational quality is limited, as the work primarily examines listening quality for differing packet loss strategies using PESQ. Interestingly, they also suggest that small adjustments to silence periods have 'almost no effect on perceived quality' without any supporting research to validate this claim. Undoubtedly, there is significant merit in a full reference objective metric that could accurately predict conversational quality, taking into account issues such as the impact of playout delays on prosody or natural turn-taking rhythm and ultimately on quality. However, whether such a metric is necessary or indeed feasible is a research question beyond the scope addressed in this paper.

In summary, significant research has examined the impact on listening quality of large scale silence adjustments and adjustments to both silence periods and active speech. None to date have addressed the impact of frequent and small playout delay adjustments (time-shifting) introduced to silence periods by Voice over IP (VoIP) jitter buffer strategies.

This gap in the literature provided the main motivation for our research, summarised and presented in this paper. This research is four-staged and structured as follows:

- Detailed subjective test carried out in May 2012 to assess the precise impact of frequent and small ($\ll$ 100 ms) silence period adjustments, typical of VoIP jitter buffer strategies, on subjective listening MOS scores (MOS-LQS).
- Comprehensive study to build on Hoene's *et al.* work in [5] and investigate the impact of such silence-period adjustments (i.e. typical of VoIP jitter buffers) on objective listening MOS scores (MOS-LQO), specifically predicted by PESQ. This research was initially presented at [14] and is more exhaustively analysed here.
- A similar and previously unpublished study on the performance of the more recent objective model POLQA.
- Comprehensive correlation analysis of both objective and subjective results.

The remainder of this paper is structured as follows. Section 2 provides background information and sets the context for our research. Section 2.1 summarises both subjective and objective approaches to speech quality measurement. Section 2.2 summarises related research. Section 2.3 outlines our research motivation and related research questions. Section 3 outlines our simulator-based approach to generating the impaired speech samples used for both objective and subjective testing. It deals with simulator details, delay profiles generated, adaptive algorithms and settings, speech samples chosen, and also summarises our speech quality assessment procedures. Section 4 presents and discusses experimental results. Section 5 concludes the paper and suggests some areas for future research arising from this paper.

2. Background

This section sets the context for our research. It firstly summarises both objective and subjective approaches to speech quality measurement. It then briefly describes related research that has touched upon similar research questions. Finally, it describes our contribution by specifying our research motivation and related research questions.

2.1. Subjective and Objective Speech Quality Assessment

Speech quality is judged by human listeners and hence it is inherently subjective. Therefore, the most reliable approach for assessing speech quality is through subjective tests. The Absolute Category Rating (ACR) test, defined by ITU-T Recommendation P.800 [15], is one of the most widely accepted methods of listening speech quality assessment. In the test, listeners express their opinions on the quality of the speech material in terms of five categories: excellent, good, fair, poor and bad with corresponding integer score: 5,4,3,2 and 1, respectively. The ratings are averaged and the result is known as Mean Opinion Score (MOS). Subjective testing is thus time-consuming, expensive and requires strict adherence to methodology to ensure applicability of results. As such, subjective testing is

impractical for frequent testing such as routine network monitoring. An interested reader can find more details about subjective testing in [16]. Arising from such limitations, objective test methods have been developed in recent years. They are machine-executable and require little human involvement. In principle, objective methods can be classified into two categories: signal-based methods and parameter-based methods. The former requires availability of speech signals to realize quality prediction process and as detailed in [17], can be further divided into two categories, intrusive or non-intrusive. Intrusive signal-based methods use two signals as the input to the measurement, namely, a reference signal and a degraded signal, which is the output of the system under test. They identify the audible distortions based on the perceptual domain representation of two signals incorporating human auditory models. Several intrusive models have been developed over recent years, like Perceptual Speech Quality Measure (PSQM) [18], Measuring Normalizing System (MNB) [19, 20], Perceptual Analysis Measurement System (PAMS) [21], Perceptual Evaluation of Speech Quality (PESQ) [22, 23] and Perceptual Objective Listening Quality Assessment (POLQA) [24, 25]. Among the models mentioned above, PSQM, PESQ and most recently, POLQA have been standardised by the ITU-T as Recommendations P.861 [26], P.862 [27] and P.863 [28] respectively. Moreover, MNB is described in Appendix II of ITU-T Rec. P.861 in order to extend the scope of the recommendation. It should be noted here that ITU-T Rec. P.861 has been withdrawn in 2001 and replaced by PESQ. In contrast to intrusive methods, the idea of the single-ended (non-intrusive) signal-based methods is to generate an artificial reference (i.e., an "ideal" undistorted signal) from the degraded speech signal. Once a reference is available, a signal comparison similar to PESQ/POLQA can then be performed. The result of this comparison can further be modified by a parametric degradation analysis and integrated into an assessment of overall quality. The most widely used non-intrusive models include Auditory Non-Intrusive QUality Estimation (ANIQUE) [29] and internationally standardized P.563 [30, 31].

Finally, parameter-based methods predict the speech quality through a computation model based on parameters rather than speech signals. The E-model is such a method, defined by ITU-T Recommendations G.107 [32] (narrowband version) and G.107.1 [33] (wideband version), and is primarily used for transmission planning purposes in narrowband and wideband telephony networks. This model includes a set of parameters, characterising end-to-end voice transmission as its input, and the output (R-value) can then be transformed into MOS-Conversational Quality Estimated (MOS-CQE) values.

2.2. Related research

To date, comparative performance analysis of per-talkspurt playout strategies to cope with network jitter (such as [1, 2, 3]) have focused on metrics such as late loss rate and average delays which are the indirect effects of such strategies, with little consideration given to either the extent or frequency of the silence period adjustments, and the impact they might directly have on quality perceived by the end user. The frequency of such adjustment is set by the talkspurt/silence ratio and thus is very much dependent on inherent speech type, but also on Voice Activity Detection (VAD) settings within VoIP applications. Such VAD settings are often user-configurable and can vary greatly across differing VoIP applications. For that reason, a speech segment identified as a silence period by one application will be listed as active speech by another. As such, for a given speech segment and network conditions, the performance of a specific adaptive strategy will be directly impacted by such settings as described by [34]. The extent of adjustments is influenced, needless to say by network conditions but also by the specific adaptive playout strategy. The qualifying phrase used by [10] that small adjustments are not noticeable is of little practical value, considering the variability in both frequency and extent of adjustments that can arise. Although no subjective listening testing to our knowledge has been done to precisely quantify this impact, some research dealing peripherally with the issue is summarised below.

Sun and Ifeachor in [7, 8] developed algorithms that seek to develop optimum buffer parameters in a trade off between delay and late packet loss. Moreover, the impact of jitter on speech quality using PESQ was investigated by Qiao *et al.* in [35] but was done by black-box testing and thus it is unclear whether the precise impact of jitter is direct (silence period adjustments) or indirect (through late packet loss). In [36], an extension to the E-model was developed to include the indirect impact of jitter, via late packet loss.

Hoene *et al.* in [5] used PESQ to investigate the impact of a single adjustment (0-1000msec) in an 8 second sample (typical of delay spikes) during both silences and active speech on speech quality. He showed that PESQ predicts significant impacts during active speech but that adjustments of up to approx. 320 ms are not noticeable during silences. In other research by Hoene *et al.* [6], he validated through subjective listening tests the behaviour of PESQ in predicting the impact of a single adjustment during active speech but these tests did not extend to similar analysis during silences.

The more extensive work of Voran [11] also deals somewhat peripherally with the issue and is summarised as follows. Voran evaluated through subjective testing, the impact of temporal discontinuities and packet loss on listening speech quality. Similar to Hoene's *et al.* work published in [6], discontinuities were applied to active speech segments only. A range of experiments were carried out to quantify the impact on MOS of such impairments. He introduced three impairments termed loss, jump and pause to speech where loss refers to conventional packet loss and was compensated for through Packet Loss Concealment (PLC), jump refers to temporal contraction of speech by dropping packets (thus without any PLC), and pause refers to temporal elongation of speech through silence inser-

tion, with PLC applied to the inserted silence. As such, the pause and jump impairments are of most interest as they involve temporal discontinuity and thus most closely reflect the type of impairment caused by per-talkspurt playout strategies. However, one key distinction between Voran's work and the operation of per-talkspurt strategies is that he applied all impairments (pause/loss/jump) to active speech segments. It is important to note that his pause impairment which introduced a silence gap within active speech was then compensated for through PLC, and his jump impairment essentially removed a segment of active speech, as if it never existed. The impact of both magnitude and frequency of each of the three impairments were examined independently as well as a combination of pause and jump. Impairments were added at random locations within G.723-encoded active speech. From [37], his main findings are summarised as follows:

- For a given frequency and magnitude of impairment, the impact of the four impairments (loss, pause, jump, pause and jump) on MOS scores was found to be roughly similar.
- As the magnitude of impairment increased, the reported MOS scores decreased at an almost linear rate. For example, at a frequency of one impairment per 100 frames, a 30/60/120 ms pause impairment resulted in the MOS score dropping by 0.21/0.41/1.15 respectively.
- As the frequency of impairment increased, the reported MOS scores decreased at a non-linear rate.

In addition to the above findings, he showed that for a very large and noticeable single adjustment (430 ms silence removal), PESQ failed to register any impact. This to some extend agrees with Hoene's *et al.* analysis in that adjustments within silences of up to approx. 320 ms are ignored by PESQ.

As emphasised earlier, both Hoene's *et al.* and Voran's detailed work introduced the impairments throughout active speech (talkspurts) only. As such, the results cannot be directly compared with per talkspurt playout strategies where the temporal adjustment impairments only occur during silences (i.e. the silence period is contracted or elongated). In particular, Voran's additional finding regarding the very noticeable 430 ms temporal adjustment (silence removal), coupled with both Hoene's *et al.* and Voran's findings that PESQ ignores such large single adjustments during silences strengthened the argument that both PESQ and POLQA need to be tested for the impact of frequent though smaller playout delay adjustments more typical of VoIP, and thus prompted us to undertake this research.

2.3. Research motivation

As outlined thus far, the literature to date has not quantified, either objectively or subjectively the precise impact of multiple small silence period adjustments, typical of VoIP applications, on quality perceived by the end user, expressed by MOS values. As described above, research by Voran and Hoene *et al.* make some contribution in this area. They both outlined firstly that the PESQ scores were not impacted by very large and noticeable single adjustments during silences. Secondly, both showed that significant adjustments during active speech did impact on subjective results (MOS-LQS) though these are quite different to silence period adjustments.

Considering all this, our primary research motivation was to address the gap in the literature by assessing the impact of frequent and small silence period adjustments on listening quality perceived by the end user, both through objective and subjective tests. In particular, we identified a number of key research questions that we wished to answer, namely:

1. What impact do frequent and small silence period adjustments have on subjective listening MOS scores?
2. What impact do frequent and small silence period adjustments have on objective listening MOS scores, specifically those predicted by both PESQ as well as POLQA?
3. Can the PESQ and/or POLQA model correctly predict the impact of frequent and small silence period adjustments on listening quality perceived by the end user, as quantified by a subjective test? If so, how accurate are those predictions?

Further questions include:

4. What relationship exists between the magnitude of adjustments and objective and subjective listening MOS scores?
5. What impact does the position of adjustments within speech samples have on objective and subjective listening MOS scores?

3. Methodology

In this section, we describe the methodology used to generate the speech samples for both objective and subjective tests, and then provide details of the testing process. A custom-built Matlab-based simulator was developed and used to generate playout adjustments (as depicted in Figure 1). The overall methodology comprised a number of stages as follows:

- Generate a series of network packet delays, consistent with varying network conditions.
- Using these delays, and simulated voice patterns (talkspurt distribution), and applied to different playout algorithms, generate a series of playout adjustments.
- Apply these set of adjustments to different locations within reference speech samples.

Using this set of degraded and reference speech samples, we carried out both ITU-T standardised subjective listening test, and objective tests, the latter using both PESQ and POLQA. This facilitated a comparison between both approaches and between PESQ and POLQA. The process of generating adjustments is described in Section 3.1. The section starts with a description of the playout adjustment simulator and ends up with a simulation work

flow and outputs. The process of applying adjustments to speech is described in Section 3.2. Details related to the actual testing are given in Section 3.3.

3.1. Playout Adjustment Generation

In order to assess, both objectively and subjectively, the impact on listening quality perceived by the end user of silence period adjustments typical of VoIP applications, a detailed simulator was built to generate such adjustments. Overall objectives were to:

- Generate a sequence of VoIP packets V with a talkspurt distribution, typical of real speech.
- Generate a range n of network delay sequences D, where each range represents different network conditions, i.e. *n*D delay values.
- For each of the n delay sequences D, generate a series of playout adjustments A that would result from applying this sequence to the VoIP packets V using typical playout algorithms.

Figure 1 depicts the playout adjustment simulator which was implemented using Matlab. The simulator was built by one of the authors for previous research as outlined in [38]. As can be seen in Figure 1, the simulator consists of three separate module blocks, namely:

- Voice Simulator Block,
- Delay Simulator Block,
- Playout Algorithm Simulator Block.

Each block is described in the following subsections.

3.1.1. Voice simulator block

The simulated voice streams were based on live speech samples. The critical factor here is the distribution of talkspurts which were extracted directly from voice tests into text files and used to reproduce speech characteristics. This was done by recording normal VoIP speech with VAD enabled and extracting the Marker bits within the RTP packet headers where '1' indicates the start of a talkspurt, and '0' represents an active speech packet within a talkspurt. An array V thus represents the distribution of talkspurt packets from normal speech – e.g. a sample subset of V [1,0,0,0,0,0,0,1,0,0,0] represents 2 separate talkspurts of duration 7 packets and 4 packets respectively.

3.1.2. Delay simulator block

Significant research has focused on modelling of Internet delay and loss characteristics [39, 40, 41, 42, 43, 44]. In [8], Sun and Ifeachor show that for VoIP, a Weibull distribution models traffic better than exponential or Pareto. Our model is designed to model the temporal relationship or burstiness of delay traffic which is commonly found and logically follows from the research that has proposed the use of bursty packet loss models. As such, we propose a series of 2-state Markov models to simulate varying network conditions. Figure 2 illustrates its application to delay modelling. The following summarises the most relevant characteristics of the models developed:

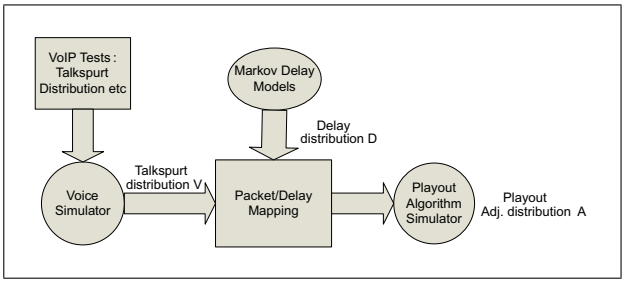

Figure 1. Playout adjustment simulator.

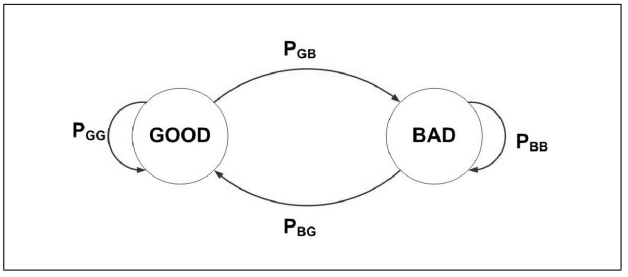

Figure 2. 2-state Markov Delay Model.

- BAD/GOOD State Jitter Level: The delay models, that were developed, used different ranges of jitter to differentiate between GOOD and BAD states. Essentially, a GOOD state had a low jitter metric (set as a % of base delays) and a multiplier was applied to this metric to represent the BAD state.
- BAD State Probability: This represents the percentage of packets that are affected by high delay variance (jitter).
- Average BAD State Burst Length: This determines how the BAD state packets are distributed. Much of the literature on network analysis has reported that both loss and delay/jitter have strong temporal dependency or burstiness. Where strong temporal dependency of jitter/delay is present, this will result in clusters of BAD state packets resulting in BAD delay/jitter bursts spanning more than one packet. Longer BAD bursts will be reflected in higher values for PBB from Figure 2.
- Using these values, we can derive values for all 4 probabilities:
  - P_{BB}: Probability that packet $n+1$ will have high jitter (BAD state) given that packet n has high jitter
  - P_{BG}: Probability that packet $n+1$ will have low jitter given that packet n has high jitter, i.e. switch states
  - P_{GG}: Probability that packet $n+1$ will have low jitter (GOOD state) given that packet n has low jitter
  - P_{GB}: Probability that packet $n+1$ will have high jitter given that packet n has low jitter, i.e. switch states

An additional important requirement from the delay block was to ensure that out-of-order packets would not arise: in reality such events are largely due to route changes and occur infrequently and thus it was important to reproduce this.

3.1.3. Playout algorithm simulator block

Two per-talkspurt adaptive strategies (namely algorithm 1 and 4 from [1] and referred to here as algorithm 1 and 2 respectively) were simulated. Both algorithms utilise linear recursive filters in tracking network conditions but differ in that algorithm 2 responds more quickly due to different parameters and also includes a spike mode that responds more rapidly to changing network conditions although it must still wait for the next silence period to do so. Both algorithms adjust playout time accordingly at the start of a talkspurt as given by

$$d_i = \alpha d_{i-1} + \big(1-\alpha\big)n_i, \tag{1}$$

$$p_i = t_i + d_i + \beta v_i. \tag{2}$$

In the above, i refers to packet i, d_i is the estimated end-to-end delay, α is the filter gain, n_i is the measured delay, p_i is the playout time, t_i is the send time, v_i is the estimated variation in delay and β is a multiplication factor. For example, in [1], the authors choose a β value of 4 for both algorithms, whereas the history factor α was set to 0.875 for algorithm 2 versus 0.998 for algorithm 1. The choice of parameters α, β and the spike detection threshold (for algorithm 2) impact greatly on the performance of these algorithms and are usually tuned to match precise network conditions i.e. from stable to unstable. For this reason, we utilised a range of values as described in the following section. As described earlier, adjusting on a per-talkspurt basis maintains the integrity of speech within talkspurts whilst altering the inter-talkspurt silence periods.

The delays from the Delay Block are mapped to the Voice talkspurt distribution series V and applied to the various playout algorithms. The delays are applied to each packet in turn and processed by the playout algorithm and adjustments made at the start of each talkspurt, indicated by a '1' in the V array. This then generates a series of playout adjustments A. As outlined in Section 2.2, the extent of adjustment is dependent not only on the network condition and playout algorithm, but also on the specific VAD settings of the VoIP application, as the latter will greatly impact on the talkspurt distribution for a given speech segment. In any event, the resulting required adjustment (silence) can be added or removed at this point before the next talkspurt is processed.

3.1.4. Simulation work flow and outputs

The overall simulator works as follows. User firstly specifies a talkspurt distribution file which is extracted from live speech and loaded as an array V. User then specifies network conditions for test. Delay block returns a sequence of network delays D corresponding to those conditions. The input parameters are:

1. Number of packets,
2. Packet interval (ms),
3. Base delay (ms),
4. BAD state burst length (ms),
5. BAD state probability (%),
6. GOOD state jitter (% of base delay),
7. BAD state jitter multiplier.

For our testing, parameters 1–4 were kept constant while parameters 5–7 were varied as outlined below to give different network characteristics, ranging from stable to unstable.

1. Number of packets = 4000,
2. Packet interval = 20 ms,
3. Base delay = 50 ms,
4. BAD state burst length = 10 Packets (200 ms),
5. BAD state probability = 20, 40, 60, 80%,
6. GOOD state jitter = 25, 50%,
7. BAD state jitter multiplier = 2, 3, 4.

Note that in arising at these parameters, particularly those relating to jitter, a detailed series of delay and jitter tests were undertaken, measuring delay/jitter between Ireland and the US/mainland Europe. Further details can be found in [38]. More recent testing by [45, 46] has highlighted the particular problem of very high jitter/delay in congested IEEE 802.11 networks. The final step was to map the delays to packets and apply them to adaptive jitter buffering (AJB) algorithms 1 and 2. This generates a series of playout adjustments A for every network test condition and playout algorithm. Figure 1 summarises the overall flow within the simulator.

In [47], details of the comprehensive tests using the simulator are presented. In summary 24 different network delay models were used, generating 24 network delay sequences D. These delay values were fed to 2 different playout algorithms as described in Section 3.1.3. For each playout algorithm, tests were repeated using different parameters such as α (history weighting - varied from 0.8 to 0.998), β (jitter multiplier - varied from 4 to 6) and spike mode threshold (only algorithm 2), see again 3.1.3 for details. Each combination resulted in a distinct set of playout adjustments A for each test scenario. Note that the voice samples V were based on 80 seconds of active speech with 40 talkspurts (Marker bit = 1) whereas the speech samples chosen for this experiment were 8 seconds long thus this also had to be factored. Essentially, a pro-rata approach was taken in that for the 80 seconds of speech used for tests, there were 40 playout adjustments so for our 8 seconds ITU-T speech samples, we implemented 4 adjustments. Arising from the full range of test combinations described above, which numbered 96, and resulting adjustments, a subset of 12 sets of playout adjustments containing 4 adjustments each were taken to represent a spectrum of network conditions ranging from a stable network to an unstable network. Table I illustrates the actual playout adjustments selected that were applied to the speech samples.

3.2. Speech samples

As normal for quality testing, 4 reference speech samples were used. The English subset of ITU-T P Supplement 23 [48] database was used for speech material, consisting of

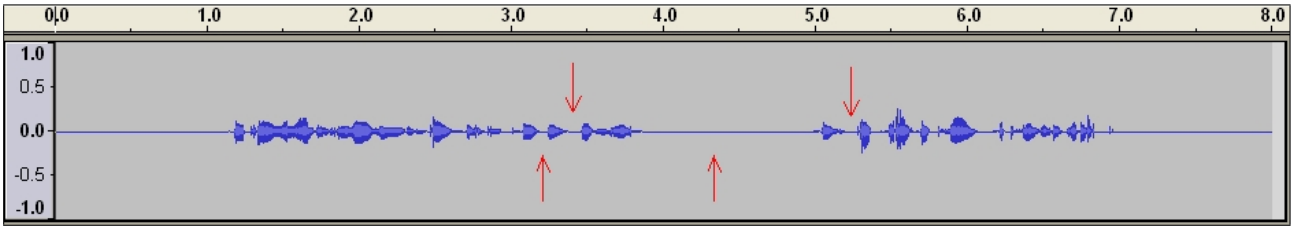

Figure 3. Demonstration of how playout adjustments were applied to 1st Female speech sample - Variant A. Arrows indicate where adjustments were placed.

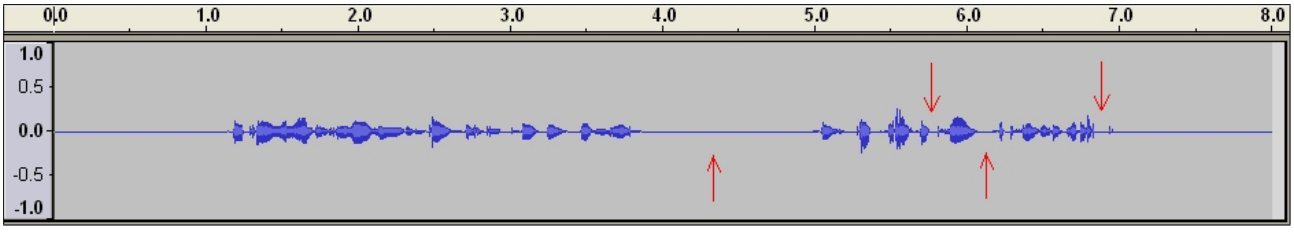

Figure 4. Demonstration of how playout adjustments were applied to 1st Female speech sample - Variant B. Arrows indicate where adjustments were placed.

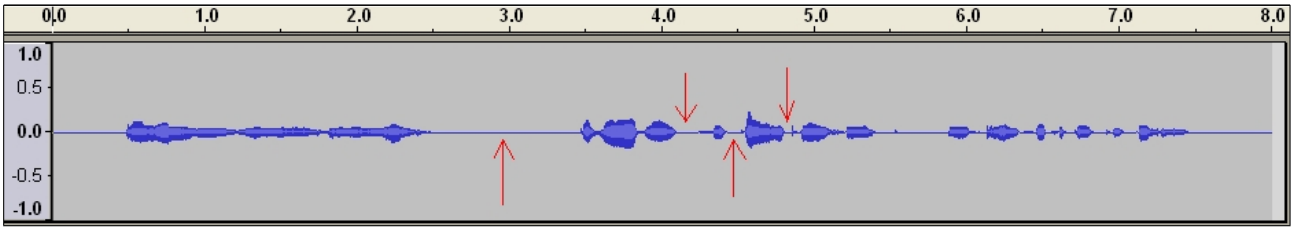

Figure 5. Demonstration of how playout adjustments were applied to 2nd Female 2 speech sample - Variant A. Arrows indicate where adjustments were placed.

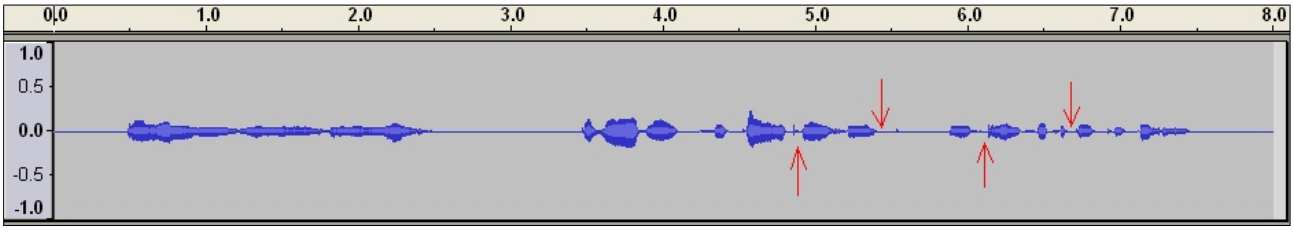

Figure 6. Demonstration of how playout adjustments were applied to 2nd Female speech sample - Variant B. Arrows indicate where adjustments were placed.

a pair of utterances with a small pause between the utterances. Two male and two female speakers uttering different sentences were included in the stimuli. The speech samples used (source samples available in the database) were 8 seconds in length and stored in 16 bit, 8000 Hz linear PCM.

Each speech sample was modified by inserting and removing silence periods to reflect the adjustments as specified above. The adjustments were a mix of positive and negative adjustments (adding and removing silence periods) as shown in Table I.

As a further experimental variable, the set of four adjustments were applied to each sample in two different locations (referred to hereafter as variant A and B). The only distinction between variant A and B is that the impairments in variant B were applied in the latter part of each sample. All the adjustments were made using a free sound editor. Figures 3–10 illustrate how the 4 playout adjustments were applied to all speech samples involved in the experiment in 2 different places (Variant A and B).

The overall result of this sampling created 96 speech samples (4 voices x 12 test conditions x 2 variants).

3.3. Speech quality assessment

The speech quality assessment process was divided into two parts, namely subjective assessment (listening test) and objective assessment (using the PESQ and POLQA models). Both assessment procedures are described in more detail below.

The ACR subjective listening test was performed in May 2012 in accordance with ITU-T Recommendation

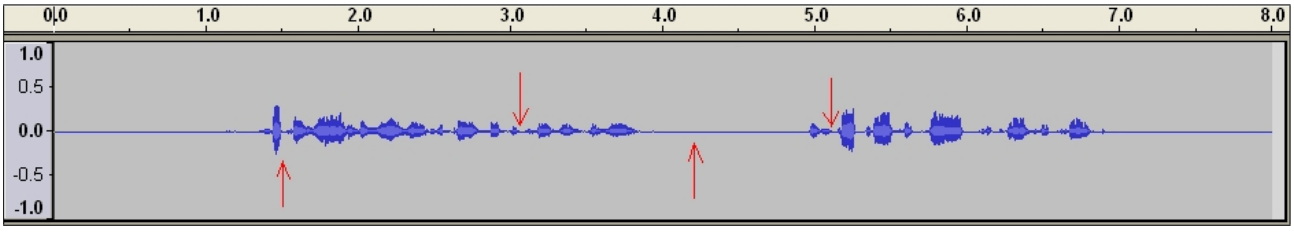

Figure 7. Demonstration of how playout adjustments were applied to 1st Male speech sample - Variant A. Arrows indicate where adjustments were placed.

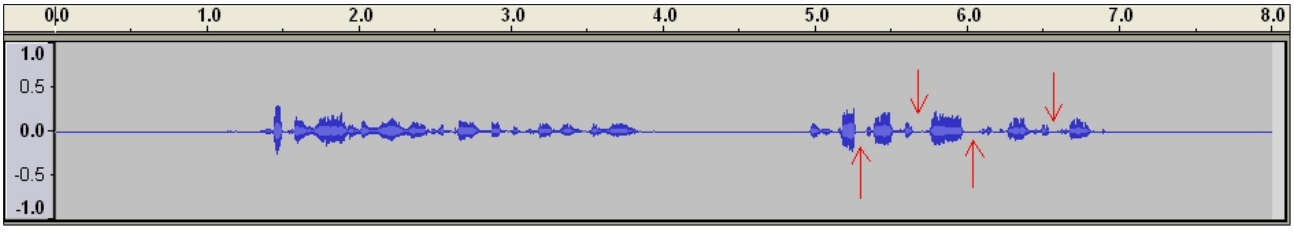

Figure 8. Demonstration of how playout adjustments were applied to 1st Male speech sample - Variant B. Arrows indicate where adjustments were placed.

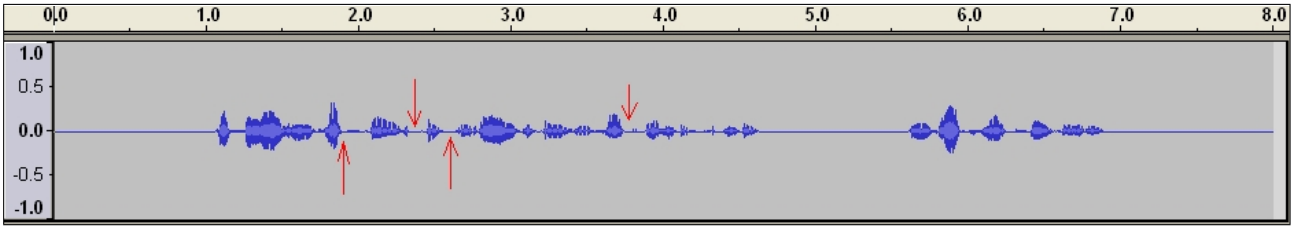

Figure 9. Demonstration of how playout adjustments were applied to 2nd Male speech sample - Variant A. Arrows indicate where adjustments were placed.

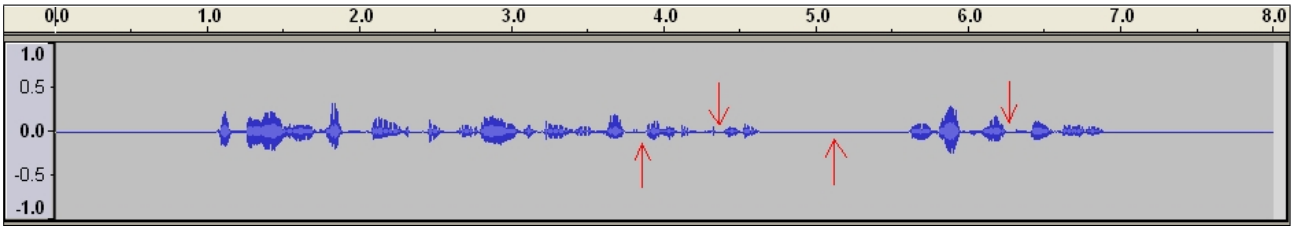

Figure 10. Demonstration of how playout adjustments were applied to 2nd Male speech sample - Variant B. Arrows indicate where adjustments were placed.

P.800 [15]. In every case, up to 2 listeners were seated in a small listening room (acoustically treated) with a background noise well below 20 dB SPL (A). All together, 30 naïve (non-expert) listeners (16 male, 14 female, 20-55 years, mean 34.43 years) participated in the test. All subjects were Irish Nationals whose first language was English. The subjects were remunerated for their efforts. The samples (96 degraded samples + 4 reference samples) were played out using high quality studio equipment in a random order and diotically presented over Sennheiser HD 455 headphones (presentation level: 73 dB SPL (A)) to the test subjects. The results of the opinion scores from 1 (bad) to 5 (excellent) were averaged to obtain MOS-Listening Quality Subjective narrowband (MOS-LQSn) values for each sample.

In the next step, the 100 samples (essentially the 96 degraded samples (using test conditions No.1-12) and 4 reference samples (Ref test condition)) were compared to their respective reference samples using both the PESQ model described in ITU-T Rec. P.862 [22, 23, 27] and the POLQA model described in P.863 [24, 25, 28], in order to get objective listening quality scores. In the case of PESQ, the output (raw PESQ scores) was converted to MOS-Listening Quality Objective narrowband (MOS-LQOn) values by the equation defined in [49].

4. Experimental results

In this section, we present experimental results for both subjective and objective assessment (PESQ and POLQA models), as well as a detailed analysis and comparison of both.

4.1. Experimental results for subjective assessment

In Figure 11, we summarise the results of subjective listening test averaged over the 4 different voices involved in the experiment, as described in Section 3. It can be seen that the impact of test conditions No. 1–12 (playout adjustments) on average MOS-LQSn scores relative to the reference samples is quite limited. On the other hand, we can also see that the subjects gave surprisingly low MOS-LQS scores to all samples, including the reference samples involved in the test i.e. the average values oscillate around 3.6 MOS. This value is quite low considering that the samples contain either no degradations (reference samples) or very moderate degradations in a narrowband context. This result warrants further analysis beyond the scope of this paper but one possibility is that the subject's opinion has been affected by their previous long-term experience with wideband telephony (wideband speech), though this was not validated. Such experience would alter their internal reference to wideband speech (extended frequency range, resulting in higher speech quality) and thus explain the lower scores given to the narrowband samples involved in the test. One other possibility relates to the fact that the range of conditions (impairments) introduced into the speech samples was quite limited. This issue is discussed in more detail in [50]. As evident, the impact of varying the extent of playout adjustments across all test conditions was very small (insignificant). However, one characteristic of note that emerged is the small impact of the location of the adjustments on scores i.e. most of the test conditions using variant B obtained slightly lower scores than the same conditions with variant A. The biggest difference (0.2 MOS) between both investigated variants (location impact) has been achieved for test condition No.10. As discussed, and evident from Figures 3–10, variant B adjustments were designed to be towards the end of the sample, so one explanation is that the distortions presented closer to the end of the sample were a bit more annoying for the subjects than those presented in the first half of the sample. In principle, this result has echoes of the so-called recency effect reported in the literature (e.g. [51, 52, 53]). However such tests used samples longer than 60 seconds. In any event, and as discussed later, the differences are not significant in the context of the confidence interval.

One three-way analysis of variance (ANOVA) test was conducted on the subjective results using test condition, voice and variant (location) as fixed factors (Table V). It should be noted here that the voice factor is a combination of voice and content. The highest F-ratio for the voice ($F = 73.12$, $p < 0.001$) was determined. The effect of voice was found to be highly statistically significant. Moreover, the test condition factor appeared to a have a weaker effect on quality than the voice factor, with $F = 0.82$, $p = 0.635$. Furthermore, the effect of test condition was not statistically significant whereas the voice factor was. The last factor investigated in the ANOVA test was the variant factor and it turns out to have a weaker effect on quality than the voice factor on its own and to not be statistically significant, similar to the test condition factor, ($F = 1.07$, $p = 0.3032$). Regarding interactions of all the involved factors, the results show that none of them is statistically significant.

Table I. Playout adjustments (in ms) applied to speech samples. Σ: Absolute sum of adjustments.

Test conditions	1st	2nd	3rd	4th	Σ
Ref	0	0	0	0	0
1	2	−2	3	−3	10
2	4	−4	−4	4	16
3	3	−3	−6	6	18
4	5	−5	−5	5	20
5	3	−6	−7	10	26
6	16	−12	−8	4	40
7	10	−17	−6	13	46
8	10	−15	−10	15	50
9	8	−23	−3	18	52
10	5	10	−30	15	60
11	−15	15	−15	15	60
12	−25	22	−8	11	66

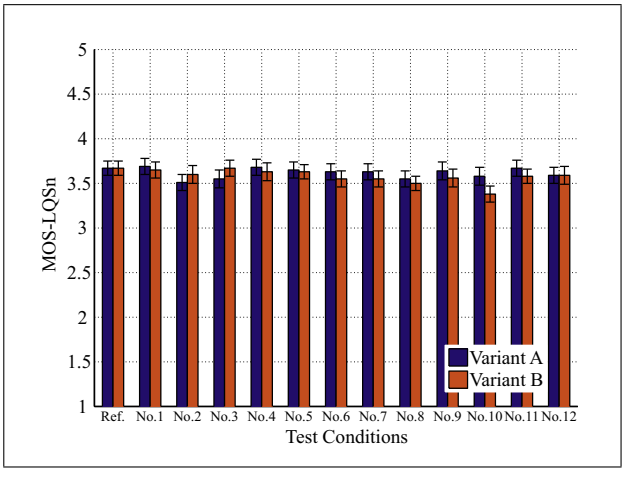

Figure 11. Effect of test conditions (see Table I for more information about the investigated test conditions) on average MOS-LQSn. The vertical bars show 95 % CI computed over 120 MOS-LQSn values (30 subjective scores per sample for 4 samples).

To summarise, the results of the ANOVA test revealed that subjects were more sensitive to the voice than to all the test conditions and variants, and no statistically significant interactions between all the investigated factors were found (assuming no impact of content due to carefully chosen speech samples). It should be noted here that a variability caused by content is considered one of sampling factors as defined in the Handbook of subjective testing practical procedures [16]. The ANOVA test also revealed that small differences between quality scores of variant A and B reported previously are not statistically significant. In other words, the results of the ANOVA test proved our assertion that the differences are not significant in the context of the confidence interval.

As is often reported in the literature, some impact relating to the voice effect was expected in this experiment but not to such an extent. A diagnostic analysis of the test data revealed that one of the voices (1st male) was liked

more than the others (i.e. over all conditions, this voice was rated on average by approx. 0.4 MOS-LQSn higher than second male voice and by approx. 0.6 MOS-LQSn higher than the female voices). It is also worth noting that both male voices have obtained higher scores than the female voices.

As can be seen above, the impact of playout adjustments is not statistically significant. This fact raises a question as to whether the impact would be higher if the Degradation Category Rating (DCR) testing approach was deployed. We considered this during the research design phase. We carried out limited DCR subjective testing (4 experts involved) and the subjects noted no degradation caused by time-shifting (playout adjustments). On this basis, we came to the conclusion that the introduced impairments would not be noticed by subjects in a DCR test. For that reason, we decided to use an ACR test.

Furthermore it should be noted that in telephony subjects have no access the speech from their conversational partner and thus ACR testing is commonly used in telephony speech quality assessment.

To the best of our knowledge, the results presented in this section are a first proof of the assertion published in the literature [1, 2, 10] that small and frequent silence period adjustments typical of VoIP playout algorithms do not have a noticeable effect on listening quality perceived by the end user.

It is interesting, at this juncture to compare our results with subjective results of Voran [11] and Hoene *et al.* [6]. Voran introduced pause/jump impairments at the rate of 1 to 4 per 3 second sentence with pause/jump magnitude of 30, 60 and 120 ms. We introduced 4 impairments in an 8 second sample, based on our observations of actual speech and using adjustments which were derived from realistic network delay models and real playout algorithms. As a result, our adjustments were typically much smaller (largest was 30 ms). Voran noted that for 1 pause/jump adjustment of 30 ms, MOS scores fell by 0.2, whereas we found no significant drop in MOS scores as the extent of adjustments increased, even for conditions 10-12 where the magnitude of some of the adjustments were similar to Voran's at 20–30 ms. One key distinction, as stated before, is that Voran applied such impairments to active speech, whereas all our adjustments were made to silence periods. Furthermore in the case of Voran's pause impairment, PLC was used. Finally, Voran reported a significant impact of a very large single adjustment (silence removal) of 430 ms. The magnitude of adjustments introduced in our samples was much smaller and as reported above, no significant subjective impact was found. Hoene *et al.* in [6] introduced very large (in comparison to adjustments introduced by jitter buffers) single adjustments into active speech and also noted a significant impact consistent with PESQ.

4.2. Experimental results for objective assessment

Figure 12 depicts the results of objective assessment done by PESQ (MOS-LQOn (PESQ)) using the same test conditions and speech samples. We can observe that the severity of the test conditions (playout adjustments) has a relatively big impact on the predicted MOS values. In summary, the MOS scores decrease as adjustments increase – i.e. as network instability increases. This is interesting in context of findings by Voran [11] that PESQ does not register any impact arising from a single 430 ms silence period adjustment (silence removal). Hoene's *et al.* results from PESQ analysis published in [5] are somewhat similar to Voran, showing no impact for single adjustments of up to approx. 320 ms. One possible explanation is that in our tests, we introduced frequent and small adjustments rather than single and large adjustments introduced by Voran/Hoene *et al.* We speculate that PESQ had difficulties with our adjustments profile during the time-alignment process. As detailed in Section 2.2, the frequency and extent of adjustments is impacted greatly by VAD settings but our test design of 4 adjustments in 8 seconds is not atypical of VoIP.

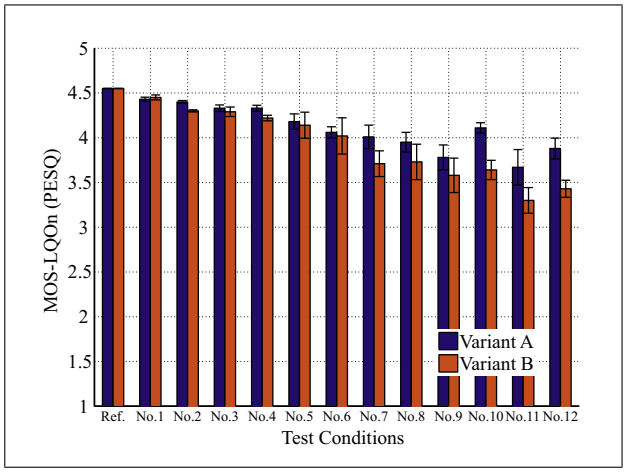

Figure 12. Effect of test conditions (see Table I for more information about the investigated test conditions) on average MOS-LQOn predicted by PESQ. The vertical bars show 95 % CI computed over 4 MOS-LQOn values (4 samples).

One positive correlation between PESQ and the subjective test results is that the biggest difference between variant A and B (0.47 MOS) was reported for test condition No.10. It should be also noted that there is a significant difference between the scores obtained for conditions No.10 and 11. This is interesting in that whilst the absolute sum of the adjustments for both conditions is the same, the individual adjustments are different. The test condition No.10 represents very different adjustments varying from 5 to −30 ms. On the other hand, the condition No.11 contains adjustments of similar magnitude; two adjustments −15 ms and two adjustments 15 ms. The second one (condition No.11) has obtained the much lower scores in both cases (Variant A and B). As such, it seems that PESQ is better able to cope with large adjustments (30 ms was largest of all conditions). In contrast, the opposite results were obtained for the subjective data i.e. listeners scored test condition No.10 lower than No.11. However, differences between the subjective results obtained for test condition No.10 and 11 are much smaller than those obtained

for objective results and of the same order as the confidence interval (see Figure 11).

As with the subjective results, lower MOS-LOQn (PESQ) scores have been reported for most of the test conditions of variant B than for the same test conditions of variant A. The average quality scores (averaged over voices) predicted by PESQ ranged from 3.665 to 4.55 MOS for variant A and from 3.3 to 4.55 MOS for variant B. It can be clearly seen in Figure 12 that the impact of adjustment location is noticeable for objective scores predicted by PESQ, especially for higher magnitudes of the investigated playout adjustments.

A diagnostic analysis of the objective data predicted by PESQ revealed that the voice impact has been much weaker than that reported above for the subjective data. In fact, intrusive signal-based models (e.g. PESQ and POLQA) are designed to focus more on impairments than on the special characteristics of voice. Due to that, such models are sometimes called impairment or degradation models. However, it seems that there is some interaction between the extent of impairments introduced in a sample (test condition) and the voice sample used, because the deviation of the MOS-LQOn (PESQ) scores between the test conditions is different for all 4 voices involved in the test. Much weaker and not statistically significant interaction of test condition and voice has been obtained for the subjective data, as shown in Table V.

Figure 13 shows the results of objective assessment done by POLQA (MOS-LQOn (POLQA)) using the same test conditions and speech samples. The trend of POLQA predictions is much more in line with the subjective results presented in Figure 11 than that of PESQ predictions. Nonetheless, it seems that POLQA was impacted more by the test conditions introducing playout adjustments with an absolute sum of adjustments above 45 ms (test conditions No. 7–12), especially those belonging to variant A. However, it is not possible to clearly identify a trend of POLQA scores, as has been done for PESQ above.

Regarding the biggest difference between variant A and B for test condition No.10 reported above for both subjective scores and objective scores predicted by PESQ, it is worth noting that this effect has not been captured by POLQA at all. In other words, the scores predicted by POLQA for variant A and B of test condition No.10 were very similar (0.02 MOS difference).

Moreover, the scores predicted by POLQA largely do exhibit same behaviour as scores obtained from subjective test from a location perspective (Variant A and B). In other words, it has been reported above that PESQ and subjects involved in the subjective test provided lower MOS scores for most of the test conditions of variant B than for the same test conditions of variant A. The average quality scores (averaged over voices) predicted by POLQA ranged from 4.10 to 4.43 MOS for variant A and from 4.08 to 4.43 MOS for variant B. This contrasts with the results for PESQ where adjustment location had a significant impact. It can be clearly seen in Figure 13 that the adjustment location plays a less important role here, except for some

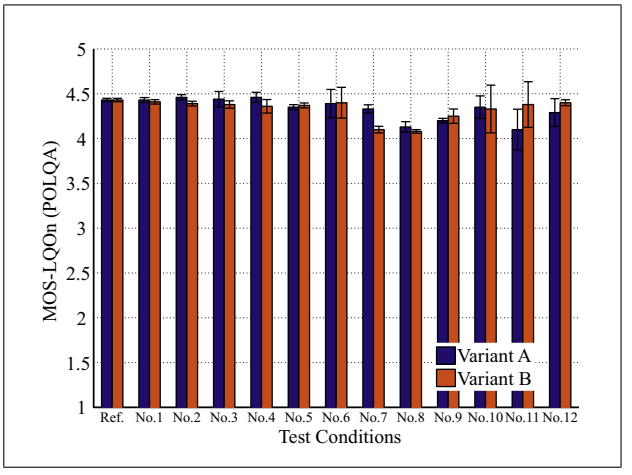

Figure 13. Effect of test conditions (see Table I for more information about the investigated test conditions) on average MOS-LQOn predicted by POLQA. The vertical bars show 95 % CI computed over 4 MOS-LQOn values (4 samples).

of the higher magnitudes of investigated playout adjustments.

A diagnostic analysis of the objective data predicted by POLQA revealed that the voice impact has been much weaker than that reported above for the subjective data. As already stated above, intrusive signal-based models (e.g. PESQ and POLQA) are designed to more focus on impairments than on special characteristics of voice. Regarding the interaction between the extent of impairments introduced in a sample (test condition) and the voice sample used reported above for PESQ model, this effect has been also obtained for POLQA model but only for female voices.

4.3. Comparison between subjective and objective quality scores

In the following subsection, subjective MOS values (MOS-LQSn) are compared to the predictions provided by both PESQ and POLQA (MOS-LQOn (PESQ/ POLQA)). The comparison is performed for all experimental conditions, i.e. all combinations of voice, test conditions and both investigated location variants. However, the MOS-LQSn values will have been influenced by the choice of conditions in the actual experiment. In order to account for such influences, model predictions are commonly transformed to a range of conditions that are part of the respective test [54]. This may be done, for example, by using a monotonic 3rd order mapping function, presuming such a function can be found.

The performance of PESQ and POLQA models is quantified in terms of the Pearson correlation coefficient R, the respective root mean square error (rmse) and epsilon-insensitive root mean square error (rmse*) as [55, 56]

$$R = \frac{\sum_{i=1}^{N}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sqrt{\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2}\sqrt{\sum_{i=1}^{N}\left(Y_i - \overline{Y}\right)^2}} \qquad (3)$$

and

$$\text{rmse} = \sqrt{\frac{1}{N-d}\sum_{i=1}^{N}\left(X_i - Y_i\right)^2}, \tag{4}$$

with X_i the subjective MOS value for stimulus i, Y_i the objective (predicted) MOS value for stimulus i, $\overline{X}$ and $\overline{Y}$ the corresponding arithmetic mean values, N the number of stimuli considered in the comparison, and d the number of degrees of freedom provided by the mapping function ($d = 4$ in the case of 3-order mapping function, $d = 1$ in the case of no regression). On the other hand, the epsilon-insensitive root mean square error can be described as

$$\text{Perror}_i = \max\left(0,\ \left|X_i - Y_i\right| - ci_{95_i}\right), \tag{5}$$

where the ci_{95_i} represents the 95% confidence interval and is defined by [56]

$$ci_{95_i} = t(0.05,\, M)\frac{\delta_i}{\sqrt{M}}, \tag{6}$$

where M denotes the number of individual subjective scores and δ_i is the standard deviation of subjective scores for stimulus i. The final epsilon-insensitive root mean square error is calculated as usual but based on the Perror with the formula (5):

$$\text{rmse}^* = \sqrt{\frac{1}{N-d}\sum_{i=1}^{N}\text{Perror}_i^2}. \tag{7}$$

The correlation R indicates the strength and the direction of a linear relationship between the subjective (auditory) and the predicted MOS values; it is largely influenced by the existence of data points at the extremities of the scales. The root mean square error (rmse) describes the spread of the data points around the linear relationship. The epsilon-insensitive root mean square error (rmse*) is a similar measure to classical rmse but rmse* considers only differences related to epsilon-wide band around the target value. The 'epsilon' is defined as the 95% confidence interval of the subjective MOS value. By definition, the uncertainty of MOS is taken into account in this evaluation. In the case of perfect agreement between subjective and objective scores, the correlation would be $R = 1.0$ and the rmse and rmse* $= 0.0$.

All R, rmse and rmse* are calculated for the raw (non-regressed) MOSn predictions and for the regressed MOS-LQOn values, obtained with the help of the monotonic mapping function (if such a function can be determined) and both (the regressed and the non-regressed MOSn predictions) are separated according to the variants, in order to get an indication of the characteristics of the PESQ/POLQA models on different types of test data.

Figure 14 compares the MOS-LQSn values with the raw model predictions (MOS-LQOn (PESQ)). The corresponding correlations R, root mean square errors (rmse) and epsilon-insensitive root mean square errors (rmse*)

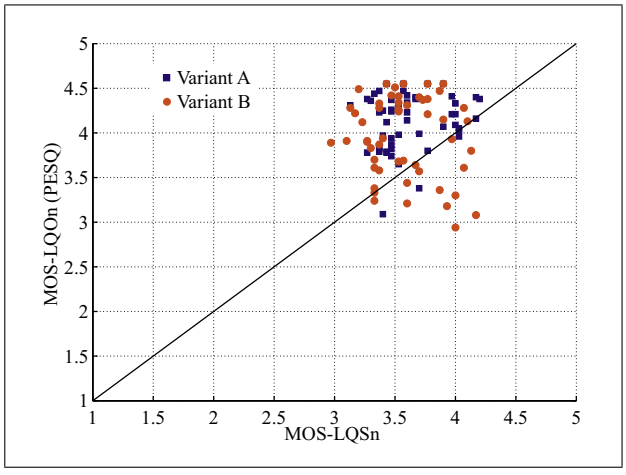

Figure 14. Subjective results (MOS-LQSn) versus MOS-LQOn (PESQ) scores (non-regressed) per sample.

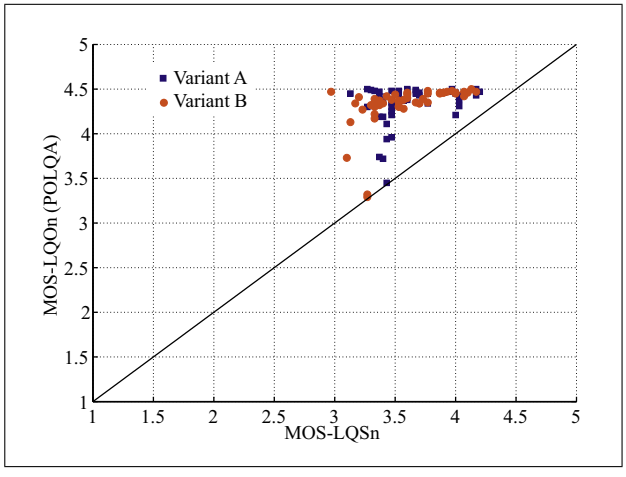

Figure 15. Subjective results (MOS-LQSn) versus MOS-LQOn (POLQA) scores (non-regressed) per sample.

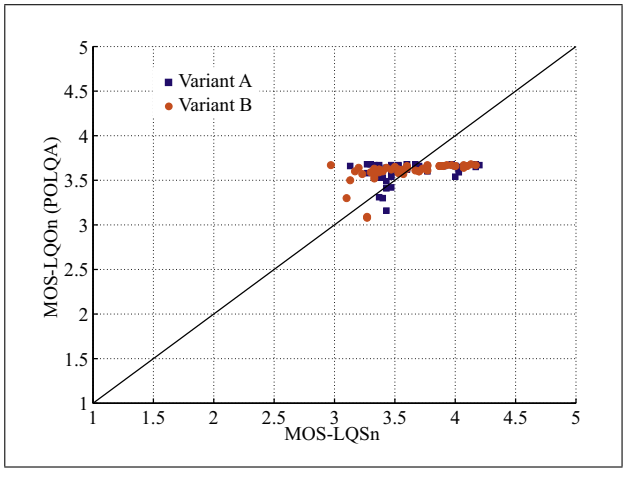

Figure 16. Subjective results (MOS-LQSn) versus MOS-LQOn (POLQA) scores (regressed) per sample.

are given in Table II. The correlation is calculated over all test conditions and voices for both adjustment locations (Variant A and B). The correlation coefficient is positive though very low ($R = 0.17$) for variant A and negative for variant B ($R = -0.15$). As normal, positive correlation

Table II. Pearson correlation coefficient, root mean square error and epsilon-insensitive root mean square error between MOS-LQSn and MOS-LQOn (PESQ) before regression.

Variant	*R*	rmse	rmse*
A	0.17	0.639	0.367
B	-0.15	0.690	0.424

Table III. Pearson correlation coefficient, root mean square error and epsilon-insensitive root mean square error between MOS-LQSn and MOS-LQOn (POLQA) before regression.

Variant	*R*	rmse	rmse*
A	0.30	0.783	0.448
B	0.47	0.806	0.475

Table IV. Pearson correlation coefficient, root mean square error and epsilon-insensitive root mean square error between MOS-LQSn and MOS-LQOn (POLQA) after regression.

Variant	*R*	rmse	rmse*
A	0.30	0.265	0.243
B	0.47	0.270	0.217

indicates that both variables increase or decrease together, whereas negative correlation indicates that as one variable increases, the other decreases, and vice versa. Moreover, the smallest rmse and rmse* were also obtained for variant A.

The 3-rd order regression as recommended in [54] leads, in this case, to a non-monotonically decreasing mapping function as opposed to a function that should be monotonically increasing. There are several options available to try to achieve monotonicity in such cases (e.g. outliers influence weighting, polynomial order change or non-polynomial function regression). In an attempt to use common polynomial regression and to avoid the sometimes questionable outlier penalization, we tried the 2-nd and 1-st order polynomial regression. The latter led to monotonic results but unfortunately the function was still monotonically decreasing. As such, we were not able to find a monotonically increasing mapping function for this data set.

Figure 17 presents the results broken down by speaker and variant. The subjective scores confirm again that there was little difference between location variants intra-speaker, significant inter-speaker variability, and very little intra-speaker variation across conditions No.1–12. As previously shown, the PESQ results exhibit a trend for higher MOS-LQOn scores in variant A over variant B, significant intra-speaker variation across conditions No.1–12, and significant inter-speaker variation.

Regarding PESQ results, i.e., very low correlation between subjective and objective data, and inability to find a monotonically increasing mapping function for this data set, we conclude that PESQ fails to correctly predict quality scores for this kind of degradation. In other words, PESQ is not able to correctly model the average user perception of the impact of frequent playout delay adjustments introduced by VoIP jitter buffers.

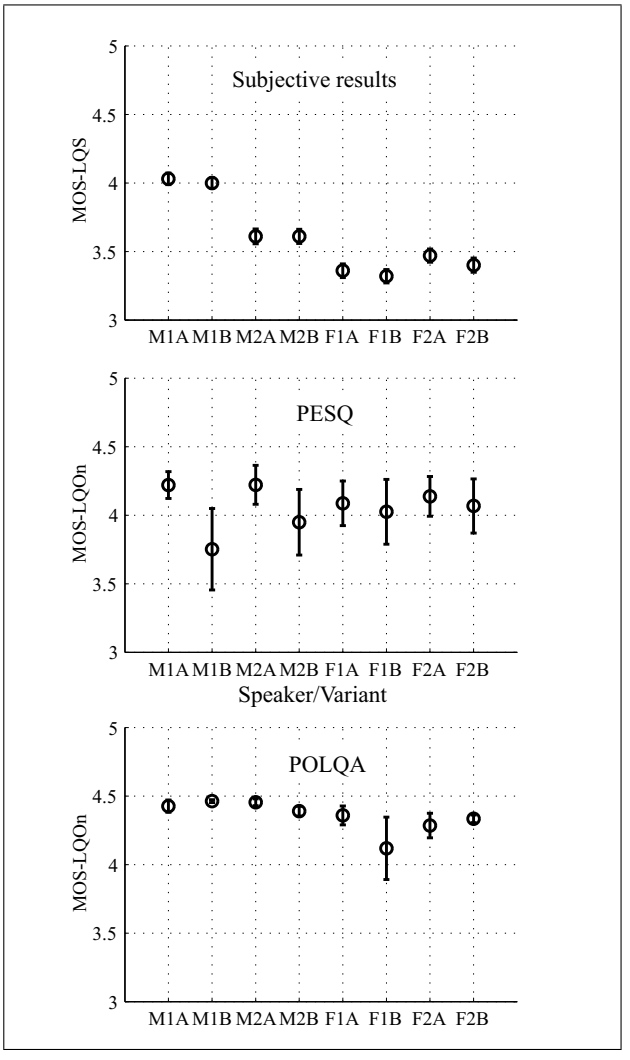

Figure 17. Dominant Experimental Factors. Results aggregated by speaker (e.g. M1 is male speaker 1) and by playout variant (i.e. A or B). The subjective scores and predictions provided by PESQ and POLQA are presented along with the 95% confidence intervals.

Moving to POLQA, results show little variation intra-speaker across variants (except for Female 1), very little intra-speaker variation across conditions No.1–12 (except Female 1), and much less variation inter-speaker. This clearly shows that POLQA performs much better than PESQ in predicting the insignificant impact of conditions No.1–12 and also the relatively insignificant impact of variants A/B (except for Female 1) Finally, the correlation data for POLQA is significantly better as shown in both Tables III and IV, and rmse data for POLQA (after regression (1st order polynomial regression applied)) is also much better than PESQ. It is worth noting that the low correlations obtained for both models are due to individual user preferences for voice.

Table V. Summary of ANOVA test conducted on the MOS-LQSn's.

Effect	SS	df	MS	F	p
Test condition (TC)	9.23	12	0.7693	0.82	0.6350
Voice	207.02	3	69.0073	73.12	0.0000
Variant	1.01	1	1.0051	1.07	0.3022
TC*Voice	18.21	36	0.5059	0.54	0.9895
TC*Variant	4.68	12	0.3899	0.41	0.9593
Voice*Variant	0.68	3	0.2282	0.24	0.8672
Error	2880.37	3052	0.9438		
Total	3121.2	3119			

We suggest a number of reasons for the particularly poor performance of the PESQ model in predicting quality scores for the investigated conditions. Firstly, we have shown that PESQ is more sensitive to the investigated adjustments than subjects are (see Figures 11 and 12), and the impact is proportional to the adjustments. Secondly, although the impact of voice on subjective scores is well known, the impact was more significant in our subjective test than expected. Thirdly, as discussed in subsection 4.1, we speculate that exposure to wideband telephony and/or the small range of impairments also influenced the subjective results and thus the prediction performance of the PESQ model. It should be also noted here that these factors may also have had an impact on the performance of POLQA model (correlation between the objective and subjective data) in this experiment.

5. Conclusions and future work

In this paper, we have investigated the impact of playout adjustments introduced by VoIP applications on quality scores obtained from a subjective listening test (MOS-LQSn) and listening quality scores predicted by both the PESQ and POLQA models (MOS-LQOn (PESQ/POLQA)). Moreover, the accuracy of both PESQ and POLQA models has also been assessed by comparing their predicted values with subjective scores. Five specific questions, outlined in Section 2.3 were addressed in our study.

Addressing the first question, we report that the impact of frequent and small silence period adjustments (playout adjustments introduced by jitter buffers in VoIP) on subjective listening quality scores is insignificant. To the best of our knowledge, the subjective results presented in this paper are a first proof of the assertion published in the literature [1, 2, 10] that the playout adjustments introduced by jitter buffers in VoIP scenarios do not have a noticeable effect on listening quality perceived by the end user.

Regarding the second question, we report that the investigated impairments (playout adjustments introduced by jitter buffers) have a significant impact on objective listening MOS scores predicted by PESQ model, whereas the impact on POLQA, though present, was much less. Note that both Voran's and Hoene's *et al.* research showed that single adjustments (430 ms in the case of Voran, 0-320 ms (approximately) in the case of Hoene *et al.*) were found to be disregarded by PESQ. Regarding question 3 and PESQ, a comparison of the subjective assessments and predictions provided by PESQ has shown that PESQ is not able to accurately predict the impact of frequent adjustments introduced by VoIP jitter buffers on listening quality perceived by the end user. Although Hoene's *et al.* research [6] has shown that PESQ model provides relatively accurate predictions for adjustments in active speech, our result suggest that PESQ performance for multiple small adjustments within silences is inaccurate. It has to be emphasized here that the PESQ model was not explicitly verified during its integration and characterization phase for frequent time shifting (playout adjustments) that results from VoIP applications with adaptive buffering over congested networks. As such our research represents a somewhat out-of-domain use case for this model.

Regarding question 3 and POLQA, our results show that POLQA is noticeably better at predicting the subjective scores. It has to be emphasized here that the POLQA model was not explicitly verified either during its design and integration phase for frequent time shifting (playout adjustments) that results from VoIP applications with adaptive buffering over congested networks. As such our research represents a somewhat out-of-domain use case for this model. It should also be noted that the POLQA results presented in this paper are also a part of characterization phase of POLQA model (and will be published in an application guide of P.863 in very limited form).

Question 4 sought to determine a relationship between the magnitude of adjustments and impact on objective and subjective listening quality scores. Results indicate an insignificant impact on subjective scores in this study. In contrast, we report a strong relationship between the extent of adjustments and objective scores predicted by PESQ. In particular, the impact of the investigated impairment increases with its extent. Regarding POLQA, we report that whilst some relationship exists, it is both much less noticeable and characterisable.

Addressing the last question, the impact of the position in the sample where adjustments are made is insignificant for subjective scores. On the other hand, the effect of the position was found to be both noticeable and consistent for objective scores predicted by PESQ model, especially for higher magnitudes of the investigated playout adjustments. Regarding POLQA however, results were much closer to subjective scores (insignificant) except for female 1.

Future work will focus on repeating this study in wideband and super-wideband telecommunication scenarios. Other questions that arise from this research and which are worthy of further investigation include:

- Why, in subjective test, did listeners return scores significantly lower than predicted PESQ/POLQA scores even for reference samples? We suggest that wideband experience and/or low impairment range plays a role.
- In our study, the extent of adjustments introduced in the samples is typical of VoIP applications experiencing moderate to severe network jitter. For VoIP applications that introduce more extreme adjustments (e.g. the impact of TCP fallback process utilised by some VoIP applications to bypass firewalls), what if any impact will this have on subjective listening quality scores?
- What precise relationship can be established between the location of adjustments and impact on subjective/objective listening quality scores? We noted that subjective scores for variant B were slightly lower though not statistically significant. Variant B adjustments were designed to be towards the end of the speech segment. We raised the point that these results suggested a recency effect, albeit with much smaller samples than used in similar tests published in the literature [51, 52, 53]. Interestingly, this relationship was more clearly evident in PESQ objective results but not so in POLQA where a consistent trend was absent.

Table V shows the results of the ANOVA test carried out on the subjective data (Dependent variable: MOS-LQSn) described in more detail in Section 4.1.

Acknowledgement

Andrew Hines thanks Google, Inc. for support.

References

[1] R. Ramjee, J. Kurose, D. Towsley, H. Schulzrinne: Adaptive playout mechanisms for packetized audio applications in wide-area networks. Proceedings of IEEE Infocom 1994, Los Alamitos, USA, June 1994, 680–688.

[2] S. Moon, J. Kurose, D. Towsley: Packet audio playout delay adjustment: performance bounds and algorithms. ACM/Springer Multimedia Systems, vol. 6, January 1998, 17–28.

[3] Y. Liang, N. Farber, B. Girod: Adaptive playout scheduling using time-scale modification in packet voice communications. Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP 2001, Salt Lake City, Utah, USA, May 2001.

[4] F. Liu, J. Kim, C. Jay Kuo: Quality enhancement of packet audio with time-scale modification. Proceedings of SPIE ITCOM 2002 Multimedia Systems and Applications, Boston, USA, July 2002.

[5] C. Hoene, H. Karl, A. Wolisz: A perceptual quality model intended for adaptive VoIP applications. International Journal Communication Systems **19** (2005) 299–316.

[6] C. Hoene, S. Wietholter, A. Wolisz: Predicting the perceptual; Service quality using a trace of VoIP packets. Proceedings of Fifth International Workshop on Quality of future Internet Services (QofIS'04), LNCS 3266, Barcelona, Spain, September 2004, 21–30.

[7] L. Sun, E. Ifeachor: Prediction of perceived converstaional speech quality and effects of playout buffer algorithms. Proceedings of International IEEE Conference on Communications (ICC 2003), Anchorage, USA, May 2003, vol.1, 1–6.

[8] L. Sun, E. Ifeachor: New models for perceived voice quality prediction and their optimization for VoIP networks. Proceedings of IEEE International Conference on Communications, Paris, France, June 2004, vol.3, 1478–1483.

[9] H. Melvin, L. Murphy: An evaluation of the potential of synchronized time to improve VoIP quality. Proceedings of International IEEE Conference on Communications (ICC 2003), Anchorage, USA, May 2003.

[10] W. Montgomery: Techniques for packet voice synchronization. IEEE Journal on Selected Areas in Communication, vol. SAC-1, no. 6, December 1983.

[11] S. Voran: Perception of temporal discontinuity impairments in coded speech - A proposal for objective estimators and some subjective test results. Proceedings of the 2nd International Conference on Measurement of Speech and Audio Quality in Networks 2003, Prague, Czech Republic, May 2003.

[12] M. Lee, J. W. McGowan, M. C. Recchione: Enabling wireless VoIP. Bell Labs Technical Journal **11** (2007) 201–215.

[13] Q. Gong, P. Kabal: Improved quality for conversational VoIP using path diversity. Proceedings of Interspeech 2011, Florence, Italy, 2011, 2549–2552.

[14] O. Stapleton, M. Melvin, P. Pocta: Quantifying the effectiveness of PESQ (Perceptual Evaluation of Speech Quality), in coping with frequent time shifting. Proceedings of the International Conference on Measurement of Speech and Audio Quality in Networks 2011, Prague, Czech Republic, June 2011.

[15] ITU-T Rec. P.800: Methods for subjective determination of transmission quality. International Telecommunication Union, Geneva, Switzerland, 1996.

[16] Handbook of practical procedures for subjective testing. International Telecommunication Union, Geneva ,Switzerland, 2011.

[17] S. Moeller, A. Raake: Telephone speech quality prediction: Towards network planning and monitoring models for modern network scenarios. Speech Communication **38** (2002) 47–75.

[18] J. G. Beerends, J. A. Stemerdink: A perceptual speech quality measure based on a psychoacoustic sound representation. J. Audio Eng. Soc. **42** (1994) 115–123.

[19] S. Voran: Objective estimation of perceived speech quality - Part I: Development of the measuring normalizing block technique. IEEE Trans. on Speech and Audio Processing **7** (1999) 371–382.

[20] S. Voran: Objective estimation of perceived speech quality - Part II: Evaluation of the measuring normalizing block technique. IEEE Trans. on Speech and Audio Processing **7** (1999) 383–390.

[21] A. W. Rix, M. P. Hollier: The perceptual analysis measurement system for robust end-to-end speech quality assessment. Proceedings of IEEE ICASSP 2000, Istanbul, Turkey, 2000, vol.3, 1515–1518.

[22] A. W. Rix, M. P. Hollier, A. P. Hekstra, J. G. Beerends: Perceptual evaluation of speech quality (PESQ) - The new ITU standard for objective measurement of perceived speech quality. Part I: Time-delay compensation. J. Audio Eng. Soc. **50** (2002) 755–764.

[23] J. G. Beerends, A. P. Hekstra, A. W. Rix, M. P. Hollier: Perceptual evaluation of speech quality (PESQ) - The new ITU standard for objective measurement of perceived speech quality. Part II; psychoacoustic model. J. Audio Eng. Soc. **50** (2002) 765–778.

[24] J. G. Beerends, C. Schmidmer, J. Berger, M. Obermann, R. Ullman, J. Pomy, M. Keyhl: Perceptual objective listening quality assessment (POLQA). The third generation ITU-T standard for end-to-end speech quality measurement. Part I: Temporal alignment. J. Audio Eng. Soc. **61** (2013) 366–384.

[25] J. G. Beerends, C. Schmidmer, J. Berger, M. Obermann, R. Ullman, J. Pomy, M. Keyhl: Perceptual objective listening quality assessment (POLQA). The third generation ITU-T standard for end-to-end speech quality measurement. Part II: Perceptual model. J. Audio Eng. Soc. **61** (2013) 385–402.

[26] ITU-T Rec. P.861: Objective quality measurement of telephone-band (300–3400 Hz) speech codecs. International Telecommunication Union, Geneva, Switzerland, 1998.

[27] ITU-T Rec. P.862: Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. International Telecommunication Union, Geneva, Switzerland, 2001.

[28] ITU-T Rec. P.863: Perceptual objective listening quality assessment. International Telecommunication Union, Geneva, Switzerland, 2011.

[29] D.-S. Kim: ANIQUE: An auditory model for single-ended speech quality estimation. IEEE Transaction on Speech and Audio Processing **13** (2005) 821–831.

[30] L. Malfait, J. Berger, M. Kastner: P.563 - The ITU-T standard for single-ended speech quality assessment. IEEE Transaction on Audio, Speech and Language Processing **14** (2006) 1924–1934.

[31] ITU-T Rec. P.563: Single-ended method for objective speech quality assessment in narrow-band telephony applications. International Telecommunication Union, Geneva, Switzerland, 2004.

[32] ITU-T Rec. G.107: The E-model: a computational model for use in transmission planning. International Telecommunication Union, Geneva, Switzerland, 2009.

[33] ITU-T Rec. G.107.1: Wideband E-model. International Telecommunication Union, Geneva, Switzerland, 2011.

[34] W. Jiang, H. Schulzrinne: Analysis of on-off patterns in VoIP and their effect on voice traffic aggregation. Proceedings of International Conference on Computer Communications and Networks (ICCCN 2000), Las Vegas, USA, October 2000.

[35] Z. Qiao, R. Venkatasubramanian, L. Sun, E. Ifeachor: A new buffer algorithm for speech quality improvement in VoIP systems. Wireless Personal Communications **45** (2008) 189–207.

[36] L. Ding, R. Goubran: Speech quality prediction in VoIP using the extended E-model. Proceedings of IEEE Globecom, San Francisco, USA, 2003, vol.7, 3974–3978.

[37] H. Melvin, L. Murphy: Exploring the extent and impact of playout adjustments within VoIP applications on the MOS. Proceedings of the International Conference on Measurement of Speech and Audio Quality in Networks 2005, Prague, Czech Republic, Jun. 2005.

[38] H. Melvin: The use of synchronised time in Voice over IP (VoIP) applications. PhD Thesis, University College Dublin, October 2004.

[39] W. Jiang, H. Schulzrinne: QoS measurement of internet real-time multimedia services. Technical report CUCS-05-99m, Columbia Univ., NY, Dec. 1999.

[40] W. Jiang, H. Schulzrinne: Modeling of packet loss and delay and their effect on real-time multimedia service quality. Proceedings of NOSSDAV 2000, Chapel Hill, USA, Jun. 2000.

[41] W. Kellerer, E. Steinbach, P. Eisert, B. Girod: A real-time internet streaming media testbed. Proceedings of the IEEE International Conference on Multimedia and Expo, Lausanne, Switzerland, Aug. 2002.

[42] R. Koodli, R. Ravikanth: One-way loss pattern sample metrics. IETF RFC 3357, Aug. 2002.

[43] C. Boutremans, J. Boudec: Adaptive joint playout buffer and fec adjustment for internet telephony. Proceedings of IEEE Infocom 2003, San Francisco, USA, Mar. 2003.

[44] Z. Li, J. Chakareski, X. Niu, Y. Zhang, W. Gu: Modeling and analysis of distortion caused by Markov-model burst packet losses in video transmission. IEEE Transactions on circuits and systems for video technology **19** (2009) 917–931.

[45] H. Melvin, P. O. Flaithearta, J. Shannon, L. B. Yuste: Time synchronisation at application level: Potential benefits, challenges and solutions. International Telecom Synchronisation Forum (ITSF) 2009,, Rome, Italy, Nov. 2009.

[46] P. O. Flaithearta, H. Melvin: E-Model based prioritization of multiple VoIP sessions over 802.11e. Proceedings of Digital Technologies 2010 conference, Žilina, Slovakia, 2010.

[47] O. Stapleton: Quantifying the effectiveness of PESQ (Perceptual Evaluation of Speech Quality), in copying with frequent time shifting. Master Thesis, NUI Galway/Regis University, Colorado, USA, 2010.

[48] ITU-T P Supplement 23: ITU-T coded-speech database. International Telecommunication Union, Geneva, Switzerland, 1998.

[49] ITU-T Rec. P.862.1: Mapping function for transforming P.862 raw result scores to MOS-LQO. International Telecommunication Union, Geneva, Switzerland, 2003.

[50] A. W. Rix: Comparison between subjective listening quality and P.862 PESQ. White paper, Psytechnics, Ltd., September 2003.

[51] I. D. C. D. S. of: the relationship between instantaneous and overall subjective speech quality for time-varying quality speech sequences: Influence of a recency effect. Source: France Telecom R&D, France (N. Chateau), International Telecommunication Union, Geneva, Switzerland, 2000.

[52] J. H. Rosenbluth: Testing the quality of connections having time varying impairments. Committee contribution T1A1.7/98-031.

[53] L. Gros, N. Chateau: Instantaneous and overall judgments for time-varying speech quality: Assessments and relationships. Acta Acustica united with Acustica **87** (2001) 367–377.

[54] A. W. Rix, J. G. Beerends, D.-S. Kim, P. Kroon, O. Ghitza: Objective assessment of speech and audio quality - Technology and applications. IEEE Transaction on Audio, Speech and Language Processing **14** (2006) 1890–1901.

[55] ITU-T Del. Contr. D.123: Proposed procedure for the evaluation of objective metrics. L. M. Ericsson (Author: Irina Cotanis), ITU-T SG 12 Meeting, Geneva, Switzerland, June 5–13, 2006.

[56] ITU-T TD12rev1: Statistical evaluation. procedure for P.OLQA v.1.0, SwissQual AG (Author: Jens Berger), ITU-T SG 12 Meeting, Geneva, Switzerland, March 10–19, 2009.